Lecture 22 - Bayesian approach; model selection

Thursday, November 23, 2023 2:31 PM

today:
- finish variational inference
- Bayesian
- model selection & causality

mean field continuation

$$\min_{q \in Q_{MF}} KL(q \| p)$$

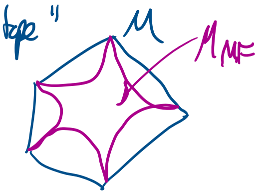
$Q_{MF} = \{ q : q(x) = \prod_i q_i(x_i) \}$

- $KL(\cdot \| p)$ is a convex fct. of q

but Q_{MF} is non-convex constraint set

e.g. Ising model $\mu_{ij} = \mu_i \mu_j$ — non-convex constraint (on moment parametrization)

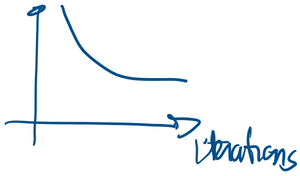
[see lecture 22 in 2017 for "marginal polytope" perspective]



[lecture 22 Fall 2017 link](#)

but can monitor progress

$$KL(q^{(t)} \| p) + cst.$$



pros & cons of variational methods vs. sampling

Variational methods:

(+) optimization based
$\Rightarrow$ often faster to run & easier to debug

(−) biased estimate

$$\mathbb{E}_{q(x)}[f(z)] \neq \mathbb{E}_p[f(z)]$$

Sampling:

(−) noisy $\Rightarrow$ harder to debug

mixing problem for chains

(+) unbiased estimate

$$\mathbb{E}\left[\mathbb{E}_{q(x)}[f(z)]\right] = \mathbb{E}_p[f(z)]$$

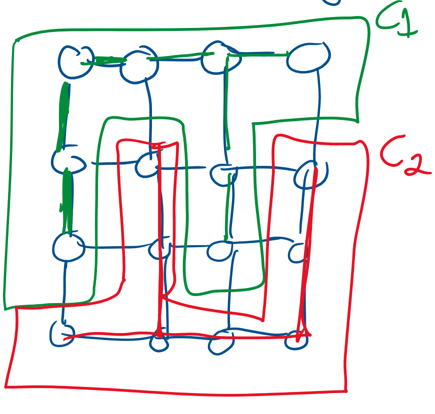
$\mathbb{E}$: with respect to random samples

$q(x)$: to make sure chain as mixed

Structured mean field

idea $q(z) = \prod_{j=1}^{K} q_j(z_{C_j})$ where $C_1, \dots, C_K$ is a partition of V

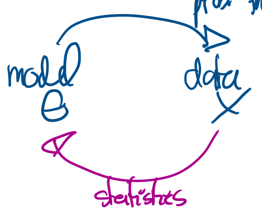
and q_j's are tractable distribution (for example tree UGM) for q_j



$$q = q_1 \cdot q_2$$

Bayesian methods

prob. theory $p(x|\theta)$



"frequentist": bag of tools $\hat{\theta}$ { MLE, reg. MLE, max entropy, moment matching, ERM, ...

"subjective Bayesian"
→ use prob. everywhere there is uncertainty

→ focus on $\underbrace{p(\theta|\text{data})}_{\text{posterior}} \propto \underbrace{p(\text{data}|\theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}$

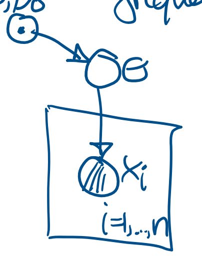
Caricature: • Bayesian is "optimist"
they think you can get "good" models
⇒ obtain a method by doing inference on model

hyperparameters for prior → α_0, β_0

• frequentist is "pessimist" → use analysis tools

Example: biased coin

$X_i | \theta \sim \text{Bernoulli}(\theta)$



e.g. $\theta \sim \text{Unif}[0,1] = \text{Beta}(1,1)$

$p(\theta) = \text{Beta}(\theta | \alpha_0, \beta_0)$

$p(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$

$$\text{posterior} \propto \left(\prod_i p(x_i|\theta)\right) p(\theta)$$

$$= \theta^{\overbrace{\sum_i x_i}^{\triangleq n_1}} (1-\theta)^{n-\sum_i x_i}\, \theta^{\alpha_0 - 1}(1-\theta)^{\beta_0 - 1}\, \mathbb{1}_{[0,1]}(\theta)$$

$$\Rightarrow p(\theta|\text{data}) = \text{Beta}(\theta | \alpha_0 + n_1, \beta_0 + n - n_1)$$

"conjugate prior" to the Bernoulli likelihood model

Conjugate priors

consider a family F of dist. on θ $\quad F = \{p(\theta|\alpha) : \alpha \in \mathcal{A}\}$

say that F is a "conjugate family" to observation model $p(x|\theta)$

if posterior $p(\theta|x,\alpha) \in F$ for any $x \sim X|\theta$

i.e. $\exists$ some $\alpha'(x,\alpha)$ s.t. $p(\theta|x,\alpha) = p(\theta|\alpha')$

sidenote: If use conjugate priors in a DGM then Gibbs sampling can be easy

[eg this is case in LDA topic model]

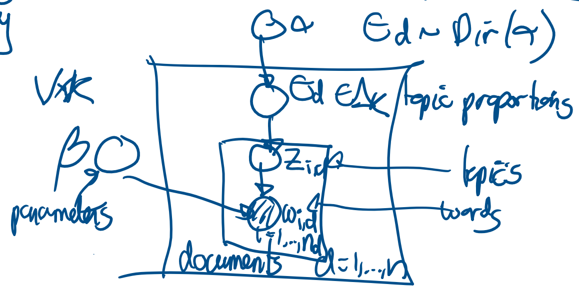
$\beta\ \alpha \quad \theta_d \sim \text{Dir}(\alpha)$

then Gibbs sampling can be easy

in LDA topic model]

$\theta_d \sim Dir(\alpha)$

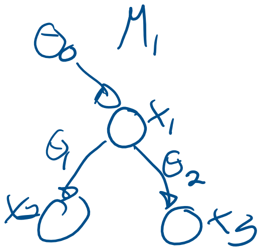
example hwk 1
Dirichlet prior is conjugate for multinomial likelihood model

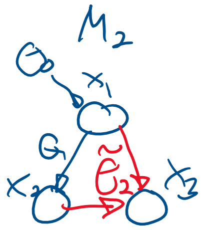


15h31

model selection

say we want to choose between 2 DGM





$p(x_3 \mid x_1; \theta_2)$ $\qquad$ $p(x_3 \mid x_1, x_2; \tilde{\theta}_2)$

[note here that "$M_1 \subseteq M_2$"] $\qquad$ "cavalier notation"

as a frequentist: $\hat{\theta}^{MLE}_{M_1} = \underset{\theta_0,\theta_1,\theta_2}{\mathrm{argmax}} \log p(\text{data} \mid \theta_0, \theta_1, \theta_2, \text{"model} = M_1\text{"})$

$\hat{\theta}^{MLE}_{M_2} = \underset{\theta_0,\theta_1,\tilde{\theta}_2}{\mathrm{argmax}} \log p(\text{data} \mid \theta_0, \theta_1, \tilde{\theta}_2, \text{"model} = M_2\text{"})$

↳ different space than θ_2

how to choose between models?

can't compare $\log p(\text{data} \mid \hat{\theta}^{MLE}_{M_1}, M=M_1)$ vs $\log p(\text{data} \mid \hat{\theta}^{MLE}_{M_2}, M=M_2)$

because LHS ≤ RHS since $M_1 \subseteq M_2$
(i.e. you would always choose "bigger model")

→ as a frequentist, use cross-validation or validation set

i.e. $\log p(\text{test data} \mid \hat{\theta}^{MLE}_{M_i}(\text{train data}), M=M_i)$

Bayesian alternatives

true Bayesian ⇒ sum over models (integrate out uncertainty about M)

introduce a prior over models $p(M)$

$$p(x_{new} \mid \underset{\text{data}}{D}) = \sum_M \underbrace{p(x_{new} \mid D, M)}\, p(M \mid D)$$

$$\sum_M \left[\int_{\theta \in \Theta_M} p(x_{new} \mid \theta_M, M)\, p(\theta_M \mid D, M)\, d\theta \right] \overbrace{p(M \mid D)}^{\text{sum over posterior over models}}$$

$$\sum_M \Big(\int_{\theta \in \Theta_M} p(x_{new} \mid \theta_M, M)\, \underbrace{p(\theta_M \mid D, M)}_{\text{posterior on } \theta \text{ given data } D \text{ \& model } M}\, d\theta \Big)\, p(M \mid D)$$

standard Bayesian predictive dist. for one model

$p(M \mid D)$: doing model averaging

⊛ in model selection, forced to pick model

$\Rightarrow$ pick model that maximizes $p(M \mid \text{data})$ or $p(\text{data} \mid M)\, p(M)$

$$\boxed{p(\text{data} \mid M) = \text{"marginal likelihood"}}$$

$$p(\text{data} \mid M) = \int_{\Theta_M} p(\text{data} \mid \theta_M, M)\, p(\theta_M \mid M)\, d\theta$$

to compare two models, look at:

$$\frac{p(M = M_1 \mid D)}{p(M = M_2 \mid D)} = \underbrace{\frac{p(D \mid M_1)}{p(D \mid M_2)}}_{\text{Bayes factor}} \; \underbrace{\frac{p(M_1)}{p(M_2)}}_{\text{prior ratio}}$$

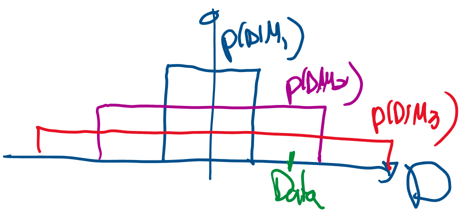
"uniform prior over models" $\Rightarrow$ then pick among k models $M_1, \dots, M_k$

by $\max\; p(\text{data} \mid M = M_i)$

"empirical Bayes" "type II ML"

when # of models is "small", then this approach is "fine" (i.e. won't overfit)

Zoubin's cartoon: suppose $M_1 \subseteq M_2 \subseteq M_3$



$p(D \mid M)$ is normalized over D

vs.

$p(D \mid \hat{\theta}_{MLE}(D), M)$ [can overfit badly]

but type II ML can still overfit if have too many models

say e.g. $p(D \mid M) = \delta(D, M)$

how to compute marginal likelihood:

use approximations
- variational inference
- sampling

use approximations — variational inference
 — sampling

simple approximation → Bayesian information criterion (BIC)

Causality:

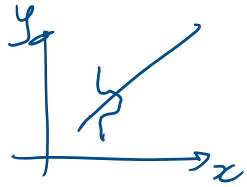
structural causal model: graph model + intervention model

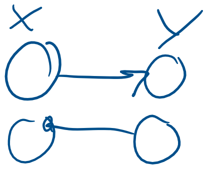
$$p(x) = \prod_{i=1}^{n} p(x_i \mid x_{\pi_i}, \theta_i)$$

semantic of intervening on node J

$$p(x \mid \text{intervention on } J) = \Big(\prod_{i \neq J} p(x_i \mid x_{\pi_i}, \theta)\Big)\, p(x_J \mid \text{intervention})$$

identify causal direction — via parametric assumptions
 — via interventions





see thoughts of Bernhard Schölkopf on causality:

https://arxiv.org/abs/1911.10500
(and references therein, e.g. his book:)

Elements of Causal Inference, 2017
By Jonas Peters, Dominik Janzing and Bernhard Schölkopf
https://mitpress.mit.edu/books/elements-causal-inference
(available for free online)